hours, and the system busy hour for the LEC is not the hour during which it receives the most terminating traffic. During some other hour (perhaps the CMRS busy hour), the LEC receives 125 units of traffic. If the ratio of total traffic to busiest hour traffic is 10 for traffic originated on the each network, total CMRS-originated traffic would be 1250 units, and total LEC-originated traffic 1000 units. Again in Case B, the LEC terminates more total traffic during a 24-hour period, but would not have to add more capacity for terminating traffic than would the CMRS provider.[14]

3. LEC-CMRS Traffic Patterns

To obtain factual information on the time profile of CMRS traffic, and on interconnected traffic between LECs and CMRS providers, the CTIA has collected data from member systems.

The collected information shows, as expected, that the amount of traffic carried on cellular systems varies throughout the business day and has a pronounced peak. A composite traffic profile for surveyed systems reporting hourly traffic patterns is shown in Figure 1.[15]

This composite traffic profile shows an overall busy hour peak for traffic from 4-5 PM. This composite result is consistent with the busy hours reported for cellular systems. A total of 51 percent of responses reported a cellular system busy hour of 4-5 PM (and an additional 20 percent reported a system busy hour of 5-6 PM). Survey information, although fragmentary on this point, suggests that the cellular system busy hour of 4-5 PM also may be the hour during which LECs deliver the most traffic for termination.[16] If this is accurate, cellular systems must terminate the largest volume of LEC-originated traffic during the cellular system busy hour.

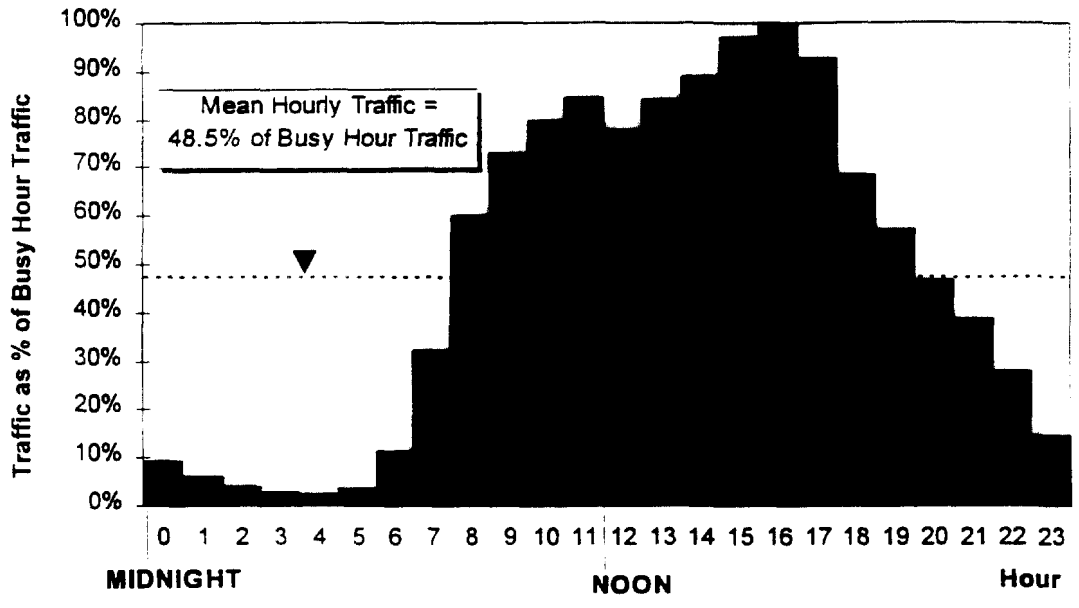
---

[14] This assumes that the amount of terminating traffic received by the LEC is not sufficiently great to shift its system busy hour.

[15] The data from which this traffic profile was calculated were for average business day traffic.

[16] One explanation for this pattern would be that the time profile both of calls both placed and received by cellular subscribers is strongly influenced by when subscribers are in their cars or have their portable units tuned on.

Figure 1
Composite Traffic Profile of Responding Cellular Systems



Source: Survey of CTIA members and CRA calculations.

This survey information suggests that the busy hours of cellular systems and LECs often will not be the same. Our understanding is that the busy hour for many LEC facilities, and often the system busy hour, is in the late morning or early afternoon, rather than the later afternoon. Only 2 percent of responses reported a cellular systems busy hour between 10 AM and noon, and only 5 percent reported a cellular system busy hour between 1 PM and 3 PM. If this is accurate, the traffic delivered to a LEC for termination would be at a maximum for many LECs outside their system busy hour, assuming, as seems likely, that the cellular system busy hour of 4-5 PM is also the hour when cellular systems deliver the most traffic to LECs for termination. LECs would receive a smaller volume of traffic for termination during their busy hour. The traffic profile in Figure 1 shows the volumes of cellular system traffic at 11 AM - noon and 2-3 PM are roughly 85 and 89 percent as large as traffic volumes during the cellular system busy hour. As in the hypothetical example in Case B above, this pattern also would make the amount of traffic

each must terminate during its system busy hour more nearly balanced than the flow of total traffic.

In the composite traffic profile of Figure 1. mean hourly traffic over the business day is slightly less than half as great as busy hour traffic. and total traffic during the business day is about 11.6 times busy hour traffic. Traffic profiles for LECs show a ratio of busy hour traffic to mean hourly traffic of about 2.5 - 3. implying that total traffic is roughly 8 - 10 times as large as busy hour traffic.[17] If the ratio of total calling to busiest hour calling is the same for LEC-originated traffic delivered to cellular systems as for all LEC calling, the pattern would be similar to that hypothesized in Case A above. This pattern would imply that the imbalance between total cellular-originated calling and LEC-originated calling would be greater than the imbalance in traffic terminated during each system's busy hour. Unfortunately, it was not possible to collect information on the time profile of LEC to mobile calling that could provide direct confirmation (or refutation) of the existence of this pattern.

In present day cellular systems, the time during which LEC subscribers can reach a CMRS handset often is limited, either by the amount of time cellular subscribers are near their cars or by the battery life of portable handsets. These factors, among others, result in an imbalance between total LEC to cellular and cellular to LEC traffic. It was possible to collect information from CTIA members only on the relative amount of traffic to and from LECs over a 24 hour period, but not on traffic received for termination during the busy hour of each network. Based on responses that provided sufficient data for the calculation, cellular systems on average received from LECs and terminated about a third as much total traffic as LECs received from CMRS providers and terminated.[18]

---

[17] Rolla E. Park, *Incremental Costs and Efficient Prices with Lumpy Capacity: The Two Product Case*, The Rand Corporation, Santa Monica, 1994.

[18] The reported figure is calculated from the means of responses to questions requesting the percent of cellular system traffic with various originating and terminating patterns. LEC-terminated traffic here does not include traffic passed on to IXCs (or traffic direct trunked to IXCs by cellular systems); LECs receive switched access payments from IXCs for such traffic , and this traffic may be less costly to terminate since end office switching and use of end office to tandem trunks is not required. Cellular-terminated traffic does include traffic from IXCs passed on by LECs since this traffic is just as costly for

As noted above, other evidence collected on traffic patterns suggests that the amount of interconnected traffic CMRS providers and LECs receive for termination in their busy hour may be less unbalanced than the flow of total traffic. Direct information on the balance of traffic during busy hours was not available, however, and indirect calculations based on limited traffic profile information that is available cannot be made with confidence. We have, however, prepared some calculations intended only to illustrate the magnitude of the adjustment to the total traffic balance that might be supported. These illustrative calculations derive the relative amounts of traffic each carrier receives for termination during its busy hour from total traffic flows under three different sets of assumptions. Each calculation begins with the assumption that total traffic terminated by the cellular system is one third as great as total traffic terminated by the LEC. The three adjustments made and the results of the calculations are as follows:

- Adjustment A: Non-coincident system busy hours for the cellular system and LECs, traffic terminated by the cellular system is at a maximum in the cellular system busy hour, but traffic terminated by the LEC in the LEC busy hour is 85 percent of the maximum hourly flow for terminated traffic. Traffic terminated by the cellular system in its busy hour would then be about 39 percent as large as traffic

---

the cellular system to terminate as LEC-originated traffic, and the cellular system does not receive switched access revenue from IXC, although the LEC does.

The mean percent of cellular system traffic in various categories, calculated from the data and estimates provided by CTIA members, is as follows:

Cellular-originated, LEC terminated	60.0%
Cellular-terminated, received from LEC (including IXC traffic)	19.5%
Cellular-originated to IXC, via the LEC	5.1%
Cellular-originated direct to IXC	11.4%
Cellular to Cellular	3.7%

The percentages do not add to 100 percent due to a small amount of unallocated traffic that was reported.

terminated in the LEC busy hour (rather than 33 percent as shown by total traffic data).[19]

- Adjustment B: Total daily terminated traffic is 11.6 times the maximum hourly terminated traffic for cellular-originated, LEC-terminated traffic and 8 times the maximum hourly terminated traffic for cellular terminated traffic. Traffic terminated by the cellular system in its busy hour would then be about 48 percent as large as traffic terminated by the LEC in its busy hour.[20]
- Adjustment C: Combines adjustments A and B. Traffic terminated by the cellular system is at a maximum in the cellular system busy hour, but traffic terminated by the LEC in its busy hour is 85 percent of the maximum hourly flow for terminated traffic, and total daily terminated traffic is 11.6 times the maximum hourly terminated traffic for LEC-terminated traffic and 8 times the hourly terminated traffic for cellular terminated traffic. Traffic terminated by the cellular system in its busy hour would then be about 57 percent as large as traffic terminated by the LEC in its busy hour.[21]

---

[19] Assume that total traffic terminated by the cellular system is 100 and total traffic terminated by the LEC is 300, and the ratio of total terminated traffic to maximum hourly terminated traffic equals 10 for traffic in both directions. The maximum traffic received in any hour for termination is 10 for the cellular system and 30 for the LEC. In adjustment A, traffic received by the LEC in its busy hour is 85% of maximum hourly terminated traffic, or under these assumptions, 25.5 (i.e., 30 x 0.85). Traffic terminated by the cellular carrier in its busy hour is 10, which is 39% of the 25.5 terminated by the LEC in its busy hour. This adjustment corresponds to Case A in the example discussed earlier.

[20] Assume again that total traffic terminated by the cellular system is 100 and total traffic terminated by the LEC is 300. Assume the ratio of total terminated traffic to maximum hourly terminated traffic is 8 for traffic terminated by the cellular system and 11.6 for traffic terminated by the LEC. This implies the maximum hourly traffic terminated by the cellular system would be 25.9 (i.e., 300/11.6), and the maximum hourly traffic terminated by the LEC would be 12.5 (i.e., 100/8). Assuming the each carrier receives the maximum amount of traffic for termination in its busy hour we obtain the result given, since 12.5 is 48% of 25.9. This adjustment corresponds to Case B in the example discussed earlier.

[21] Begin with the figures in the previous footnote: Maximum hourly traffic terminated by the LEC is 25.9 and maximum hourly traffic terminated by the cellular system is 12.5. Making the further adjustment that traffic terminated by the LEC in its busy hour is 85% of the maximum hourly flow of terminating traffic, the LEC terminates 22.0 units of traffic in its busy hour (25.9 x 0.85); 12.5, the traffic terminated by the cellular system in its busy hour, is 57% of 22.0.

These calculations are no more than illustrative (although each is at least suggested by available information). They do, however, indicate that the balance of total traffic exchanged could be quite different from the balance of traffic imposing capacity costs on the terminating carriers. Starting with total LEC-terminated traffic that is 3 times cellular-terminated traffic, the adjustments reduce LEC-terminated traffic to as little as 1.8 times cellular-terminated traffic. Even such adjusted figures for the balance of traffic tell only part of the story of the balance of costs imposed by interconnection. Those costs depend on the level of capacity cost per minute in each carrier's busy hour as well as the balance of traffic that imposes capacity costs. Before turning to this issue, however, it is important to remember that all the traffic data discussed above are for cellular systems, and reflect current technology and features of cellular system, the current pricing of cellular systems, and the current level of interconnection payments made and received (or not received) by cellular systems.

The next generation of CMRS systems will likely include advances in technology, service features, and pricing options designed to increase traffic per subscriber. Low-power digital handsets, extended battery life, and the capability of receiving and displaying caller number identification will encourage subscribers to use portable terminals throughout the day. Integration of a mobile telephone number with voice messaging will enable subscribers to return calls in instances when they cannot be reached directly. Pricing innovations, such as the free first minute for received calls promoted by the first operating PCS system, can both stimulate total traffic and increase the fraction of minutes originated on the CMRS system. Overall, as CMRS handsets become increasingly good substitutes for fixed telephones, the future traffic patterns of CMRS systems are likely to more closely resemble those of wireline local telephone systems, with the result that the total flow of traffic terminated by LECs and by CMRS systems will come to be more nearly balanced.

An early report lends some support to the proposition that the flow of traffic exchanged will become more balanced between CMRS providers and LECs. The first

welfare properties of these compensation arrangements. The following section discusses the effects of these arrangements on transactions costs. Section VII discusses how compensation arrangements may affect the development of competition and dynamic efficiency.

V. The Efficiency of Price Signals

Prices shape purchasing behavior. Lower prices encourage purchases and higher prices discourage purchases. The level of demand for various products or services in turn directs the allocation of resources and determine how much of which products and services are produced. The policy concern is that the structure and level of prices be set so that they can perform this allocative function efficiently. Prices perform their allocative function most efficiently when their structure and level of prices for a service accurately signal to purchasers the costs of producing that service. It is this function of prices that leads to the prescription, in standard textbook models, that for maximum efficiency price should equal marginal cost.

In this section we discuss how good a job the prices implied by usage sensitive and bill and keep arrangements are likely to do in providing signals that will induce efficient choices by consumers. It may seem obvious that usage sensitive pricing will perform better in this comparison. The simple case against bill and keep is easily stated: Bill and keep arrangements set a price of zero on additional traffic delivered to another network for termination,[23] while most costs of terminating traffic are usage sensitive. Therefore, the simple case concludes, a price of zero sends an inefficient signal since consumers will make additional calls without taking into account the cost imposed by additional traffic. Instead, the simple case suggests that usage sensitive costs should be recovered with usage sensitive prices; price then reflects the cost of additional usage, and will send efficient signals to consumers and the marketplace.

---

[23] As seen above, however, this does not mean that interconnection services taken as a whole are free under bill and keep. Under bill and keep, CMRS providers and LECs each must incur a cost in exchange for receiving interconnection services.

But this argument is too simple. First, it ignores the effects of compensation arrangements on total costs and on dynamic efficiency. Second, a full analysis of the static efficiency of pricing signals is both more complicated and less clearly favorable to usage sensitive pricing than is admitted by this argument. A full analysis should consider both the actual structure of costs as well as the structure of pricing that will be achievable in practice. The efficiency of pricing signals depends on having the structure of prices match the structure of costs, not merely having the average level of prices matching the average level of costs. To begin this analysis, the next section looks at the structure of interconnection costs.

A. The Structure of Interconnection Costs

Interconnection and the exchange of traffic involves at least two kinds of facilities and costs that should be distinguished. Each has its own structure that should be considered in designing prices to recover that category of cost:

1. Costs of facilities dedicated to interconnected traffic. The leading example is the cost of trunks connecting the networks.
2. Costs of the network facilities that each provider uses both to terminate interconnected traffic and to carry and terminate other traffic.

We discuss briefly the structure of the first of these types of cost, and the appropriate structure of prices to recover these costs. We then look in more detail at the cost structure for shared network facilities; these are the interconnection costs most often thought of as usage sensitive.[24]

1. Costs of Dedicated Facilities

The cost of the dedicated circuits connecting CMRS and LEC networks depends on the number and characteristics of the circuits installed, and only indirectly on the amount of traffic carried over those circuits. Costs are driven by the amount of circuit

---

[24] A third category of possible costs is one-time costs of adapting CMRS or LEC networks to handle or monitor interconnected traffic. Clearly there will be inefficiencies in recovering one-time costs with continuing charges on usage.

capacity in place. Changes in traffic may change the capacity needed. but traffic may also change without affecting these capacity costs if the change in traffic can be accommodated by the capacity already in place. Because these costs do not vary directly with traffic, it will not be efficient to recover them with a simple charge on all units of traffic. As with charges for private lines, and for the same reason, charges to recover these costs should be structured to depend on circuit capacity, not the volume of traffic carried.

The rule for efficient pricing is simple if separate circuits are dedicated full time to carry LEC to CMRS traffic and CMRS to LEC traffic. In this case, the LEC and CMRS provider should each be responsible for the cost of the trunk capacity carrying the traffic it originates. We understand, however, that traffic in both directions often shares the same circuit capacity. The volume of traffic in each direction might then be used to share the cost of this shared capacity, but it will not be efficient to accomplish this with a simple usage charge. First, it will be more efficient to base the sharing of a cost that depends on circuit capacity on relative usage, than to set a per unit usage charge that causes the total amount paid to fluctuate with total usage rather than circuit capacity. Second, it will be more efficient for the sharing of costs to depend on the circuit busy hour usage than on total usage, since it is busy hour usage that will drive the capacity needed.

Finally, it may be efficient to use sharing rules rather than traffic measurements to determine the division of capacity costs. One such rule, often used for trunks interconnecting adjacent LECs, is for each carrier to bear the full cost of the trunks up to some defined "meet point" midway between the networks. Such a rule has the virtue of causing the cost borne by each carrier to vary with the amount of installed circuit capacity, while still potentially saving costs of monitoring usage over the trunks and billing for those costs.

2. Shared Terminating Network Costs

An interconnected CMRS or LEC network terminates traffic originated by subscribers to the other network and directed to its own subscribers. Terminating traffic

from the interconnected network is mingled with other traffic carried on the terminating network, sharing use of the same switch and trunk facilities, and (in the case of CMRS networks) of cellsite and associated equipment used to establish and maintain radio connections with subscribers. Terminating the traffic imposes a cost on the carrier because an increase in the amount of terminating traffic, like an increase in other traffic carried by the same facilities, can increase the needed capacity.

The costs imposed by terminating traffic are fundamentally costs of increasing capacity, just as the costs of the interconnecting trunk facilities are costs of providing the necessary capacity. The difference is that the capacity of an interconnecting trunk is dedicated to interconnection service, so the cost of that trunk can be identified as a cost of interconnection. Where interconnected terminating traffic shares use of network facilities with other traffic, no identifiable facilities are dedicated to interconnected traffic in general, or to terminating traffic in particular.

Still, the fact that these are costs of capacity determines the structure of shared network costs. Only additional traffic that presses on the capacity of network facilities imposes a cost. Since facilities are sized to provide a specified grade of service during the busy hour, only increases in traffic during the busy hour require investments to increase capacity. It is accurate to say that the costs of the shared network facilities are usage sensitive, but only in the sense that they vary with *some* usage, namely usage during the busy hour. These costs are not sensitive to, or increased by, all increases in traffic. Additional traffic outside the busy hour of a facility, which can be accommodated without increasing capacity, imposes almost no additional costs.

Two further complications in the structure of these costs are relevant. First, it is a simplification to talk only of the system busy hour for the entire network. Different facilities or components of the network can have different busy hours. For example, many portions of local exchange networks carry the most traffic and have their busy hour during the middle of the day. However, the busy hour is in the early evening for some end office switches in residential areas and for the common transport trunks to some residential end offices. The second complication is that the costs of adding capacity to a particular type of facility may vary with the geographic location of the facility, or perhaps

the type of equipment at particular locations. These two complications mean that the cost imposed by a minute of terminating traffic does not depend only on whether it occurs in "the" busy hour. The routing of a call will determine how many of the facilities used to terminate that call have their busy hour at that time, and the costs of adding capacity to those particular facilities.

We now turn to the implications of this cost structure for efficient pricing.

B. Matching the Structure of Prices and Costs

The review above shows the basic flaw in the simple argument in favor of usage sensitive pricing. Shared network costs may be sensitive to particular traffic flows, but it does not follow that a uniform price on usage accurately sends a signal of underlying costs. Not all minutes of usage will impose the same costs. This section analyzes in more detail the static efficiency of pricing signals from usage sensitive prices and from bill and keep arrangements. Prices can be usage sensitive, of course, without being based on cost. Any claim that usage sensitive prices send efficient signals of costs, however, will depend on their being based on costs. Therefore the discussion below only analyzes usage sensitive prices based on cost. The first step in this analysis, then, is to see how prices would be derived from cost.

1. Derivation of Cost-based Prices

The following is a very simplified description of how prices for a particular service would be derived from the costs of the set of facilities and related expenses that would provide the capacity necessary to provide that service. The discussion focuses on only a few key steps in the process that are used in the discussion below, and abstracts from many important issues that must faced in deriving unit costs and prices.[25]

---

[25] Among the issues not considered are whether the costs being measured (and on which prices are to be based) are long run or short run costs, and are marginal or service incremental costs. Another issue not considered is the appropriate way of determining the amount of traffic from which the cost will be recovered when capacity is lumpy and more capacity is installed than is immediately needed.

First, determine the costs of the facilities that are used. In the case, the facilities would include various trunks and switches. The investment to create the capacity provided by these facilities include both equipment and installation costs.

Second, the cost of the capacity must be converted to a cost per unit of time. These facilities are long-lived, and their costs will be recovered over their life. Using a depreciation rate and a discount rate, the investment cost is converted to an equivalent cost per unit of time, for example, an annualized cost.

Third, expenses directly associated with operating this capacity are added to arrive at a total cost. For example, annual maintenance costs would be added to the annualized investment cost of the facilities. These steps result in a cost per unit of time.

Fourth, cost per unit time is converted to price by dividing by the number of units of billed usage of this capacity during the unit of time for which costs were calculated.

To give an example, assume that steps 1 through 3 have yielded an total annualized cost of $1 million for the CMRS capacity used both to terminate interconnected traffic originated by LEC subscribers and to carry all other traffic of the CMRS provider. The objective is calculate a price, based on this cost, that will be charged, on completed calls, for each minute of originating usage and each minute of terminating usage of the CMRS network. Say that in a year there are 100 million minutes of originating plus terminating usage; this includes all traffic using these network facilities, not just the termination of interconnected traffic. Dividing $1 million by 100 million minutes of usage yields a price per minute of originating or terminating usage of 1¢ per minute.

If, instead, only usage during a peak pricing period were to be billed, price would be calculated using total usage during the peak period. Say that annual usage during a peak period of 8 AM to 8 PM totaled 50 million minutes. The price per peak period minute would be $1 million divided by 50 million minutes, or 2¢/minute.

2. "Optimal" Pricing

Given that most costs are costs of capacity, what prices would send "optimal," efficient pricing signals? "Optimal" is put in quotation marks because this discussion

considers only the effects of pricing signals on static efficiency. Other effects on efficiency, such as the impact on costs of monitoring and billing for usage and on dynamic efficiency and competition, are ignored at this point. We refer to "optimal" pricing throughout this section for convenience, although this pricing is optimized for only one of several relevant criteria.

What prices are optimal in the static sense of sending efficient signals is influenced by both the structure of capacity costs and by the distribution of the demand for calling through the day.[26] Busy hour traffic determines the sizing of network facilities. A first cut at matching price to cost would be to set a price only for usage during the busy hour, while charging a price of zero for all other usage. A price charged only for busy hour usage would be relatively high since it would be paid on only a fraction of total usage. In the hypothetical example above, capacity costs were $1 million, and there were 100 million minutes of total traffic. If we assume the ratio for total minutes to busy hour minutes is 10 to 1, total busy hour traffic is 10 million minutes. A price on only busy hour traffic would be 10¢ a minute, ten times higher than the price of 1¢ that would be charged on all usage, since busy hour traffic is only 1/10 of total traffic.

A price applied only to busy hour usage still may not be theoretically optimal. The relatively high price will depress usage during the single hour it is charged, which may result in some other hour becoming the busy hour. This phenomenon is referred to as "peak shifting." Figure 2 illustrates the point. Panel A graphs the (hypothetical) distribution of traffic throughout a business day; the prices listed for each hour across the top of the graph show that this is the distribution of calling when the price for usage is zero at all times. The busy hour of this distribution is at 2 PM, but traffic is about 90 percent as high during several other hours. Panel B shows the effect of setting a price only for usage during 2-3 PM. Usage declines in that hour while increasing somewhat at

---

[26] For discussions of optimal pricing (in the sense used here), see R. E. Park and Bridger M. Mitchell "Optimal Peak-load Pricing for Local Telephone Calls", RAND, R-3401-1, March 1987, and Bridger M. Mitchell and Ingo Vogelsang, *Telecommunications Pricing: Theory and Practice*, Cambridge University Press, Cambridge, 1991, and the references cited in these works.

other times. The new peak is at 1-2 PM, but traffic is very nearly as high at 10-11 AM, which is a secondary peak of the distribution. Setting price this high for a single hour is not optimal, both because there is no charge for what becomes the busy hour when the peak has shifted, and because traffic has been depressed below capacity during the original busy hour. This means that the high price deters some calling that would not impose a cost.

Further pricing adjustments are required for optimality, and the direction of the needed changes is clear. Price should be somewhat lower during the original busy hour of 2-3 PM, and a non-zero price should be set for traffic during what would become the new busy hour. Charging for usage only during the original peak, and the new peak of 1-2 PM however, could further shift traffic, and create yet another peak.

The theoretical solution is to set non-zero prices for several hours, but to set a *different* price for each of these hours.[27] The price to set for each hour depends on both the underlying demand for usage at various times (here manifested by the call distribution when price is zero), and how increased prices at one time will cause usage to shift to other times. Panel C show what such a set of optimal prices might look like, and the resulting distribution of calling. In panel C, non-zero prices are charged for four hours: 10 AM - noon, and 1-3 PM, with prices ranging from 1¢/minute from 11 AM to noon to 5¢/minute from 2-3 PM. Notice that the result of these prices is to make usage the same during each of these four hours; optimal pricing smooths peak usage to create a group of busiest hours in place of a single busy hour.[28] For the other 20 hours of the day outside these busiest hours, price is set at zero.

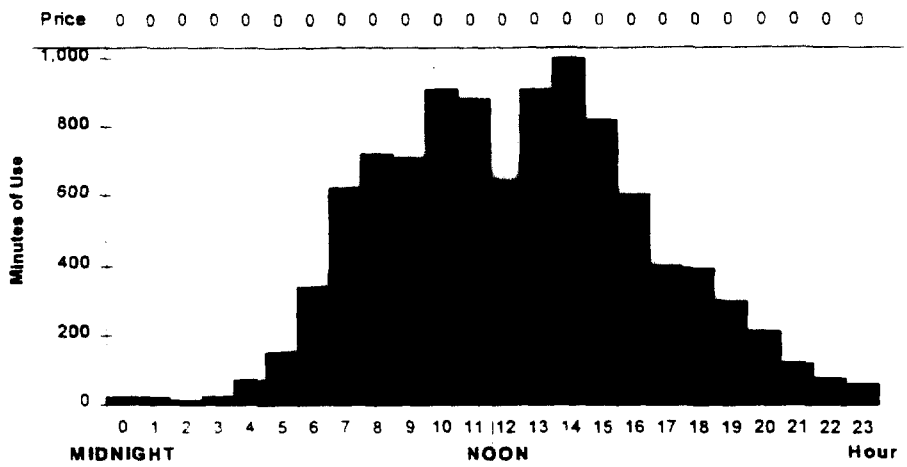
---

[27] Marcel Boiteux, "La Tarification des Démandes en Pointe", *Revue Générale de l'Electricité* 58: 321-40, 1949.

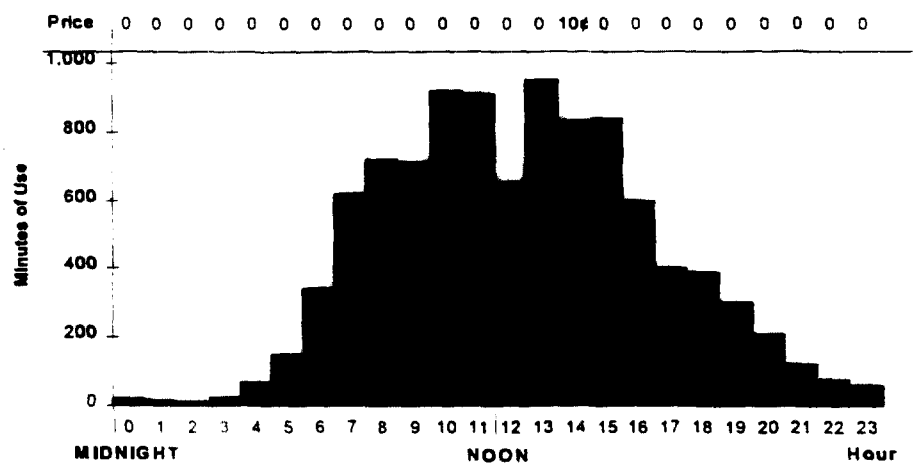
[28] The other characteristic of the set of optimal prices is that the sum of the prices should equal the marginal cost of a unit of capacity.

Figure 2
Hypothetical Traffic Profiles and Pricing

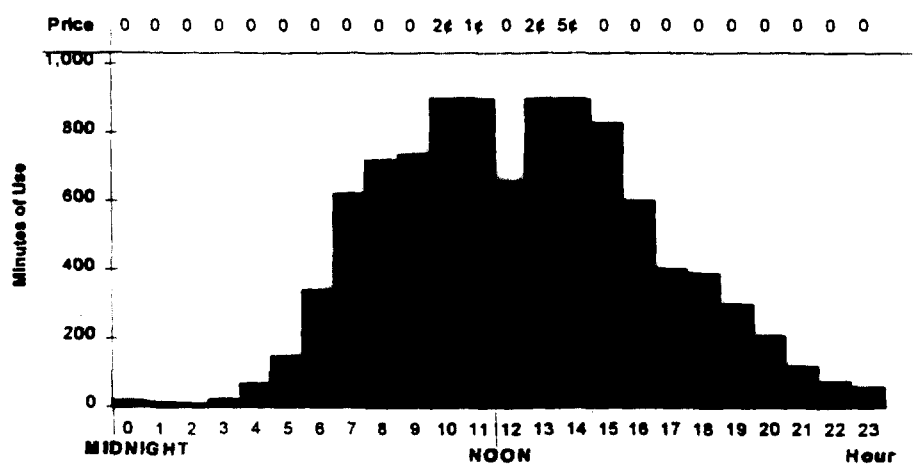
Panel A: No Usage Charge



Panel B: Busy Hour Charge Only



Panel C: "Optimal" Pricing



We can use this "optimal" price structure as a benchmark for comparing the efficiency of pricing signals send by usage sensitive pricing and bill and keep arrangements. Before turning to this, one final point is important. "Optimal" prices have been derived by considering the effects of prices on consumer demand -- that is, on the volume of traffic. Interconnection arrangements, however, set the *wholesale* price paid by the other carrier, not a *retail* price paid by consumers. An additional linkage is needed to apply these results to pricing wholesale service: retail pricing must reflect the structure and level of the wholesale price structure. There are market forces that push to create precisely this linkage. Competition pushes firms to set retail prices based on the level and structure of their costs, including the structure and level of wholesale prices they pay for various inputs. At the same time, retail prices may only approximate the structure of underlying costs, even for competitive firms. Retail prices that more accurately match costs may not occur either because trying to set and collect such prices would increase costs, or because consumers are confused by or otherwise dislike dealing with such complicated pricing. The relationship between wholesale and retail pricing and its significance are discussed in more detail later.

3. Uniform Price per Minute Compared to Bill and Keep

The point of departure for this comparison is that neither a uniform price per minute, nor bill and keep arrangements send pricing signals that are "optimal." This is a comparison of two "suboptimal" pricing structures.

Uniform Price Per Minute

A uniform price per minute never sends quite the right price signal, except by chance. All additional traffic is charged a price, even when network facilities have excess capacity, whereas the correct price signal at such times is zero since additional traffic imposes essentially no additional network costs. Uniform prices also send inefficient signals at all or most times when additional usage does impose capacity costs, because in

these periods a uniform price will be below the costs imposed by additional traffic and will send inefficient signals.[29]

Because calling distributions are uneven, with one or perhaps two pronounced peaks during the day, the optimal pricing structure would set non-zero prices for only a few hours when traffic is at or near its peak. Prices would be relatively high for most of this time, since prices charged for only this subset of traffic would recover almost all costs of increasing capacity. Compared with this optimal price structure the uniform price per minute will be lower since it will be calculated by dividing capacity costs by total traffic. Optimal prices will vary by time, and some of the lower, non-zero prices might be about the same level as a uniform price, but that would occur only by chance and likely for only a small portion of the day.

A uniform price per minute might be correct "on average," in the sense that average revenue per minute might be about the same as with optimal prices. This is a case, however, where being right on average means being wrong almost all of the time. The uniform price per minute is nearly always too high or too low, and both deviations create inefficiencies. Charging too high a price inefficiently discourages use: consumers fail to make some calls that would benefit them, even though those calls would impose virtually no costs. Charging too low a price inefficiently encourages use: consumers make calls they value less than the costs of making them. The economic term "deadweight losses" is given to reductions in welfare from prices that are too high and too low. Uniform prices will generate deadweight losses for most traffic.

Bill and Keep

Bill and keep sets a price of zero for sending additional traffic for termination. This is the optimal price and generates no deadweight loss for traffic that imposes no capacity costs. A very large part of traffic outside the system busy hour will impose no capacity costs, and much of the rest will impose only minimal capacity costs.[30] This

---

[29] Optimal pricing that varied by time would smooth the peak calling, and thus there would be more than a single busy hour when additional calling would impose capacity costs.

[30] Some traffic outside the system busy hour may impose some capacity costs because not all network facilities experience their busy hour during the system busy hour. Additional traffic at times when

means that bill and keep's price of zero is the optimal price for a very substantial portion of total traffic, and a near-optimal price for other non-busy hour traffic. In the composite traffic profile for cellular systems presented in Figure 1, over 91 percent of total traffic during the business day is carried outside the cellular system busy hour; the proportion of total traffic outside the busy hour presumably is even larger over the entire week or month.[31] For the pricing of termination service by the LEC it really is the amount of CMRS traffic terminated by the LEC during the LEC busy hour that is relevant. The proportion of cellular traffic that is delivered for termination outside the LEC busy hour could well be still larger than the proportion calculated with reference to the cellular system busy hour.[32]

Bill and keep, however, does not send optimally efficient pricing signals for all of the interconnected traffic. The bill and keep price of zero is too low during the busy hour or, more generally, for traffic that does impose capacity costs on the terminating carrier. A uniform price per minute also is too low a price for busy hour traffic. Nevertheless, since a uniform price is higher than the zero, it will be closer to the optimal price and generate smaller deadweight losses for this traffic at these busiest time of usage than bill and keep.

In sum, neither a uniform price per minute nor bill and keep always send optimal pricing signals. A uniform price will almost always be either too high or too low. Bill and keep's price of zero will send the right signals for what likely is a substantial

---

no facilities have their busy hour will impose no capacity costs. Additional traffic at times outside the system busy hour, but when some individual network facilities used by the interconnected traffic have busy hours, will impose some capacity costs. The capacity cost per minute for such traffic, however, will be very low relative to capacity cost per minute in the system busy hour so long as the facilities with their busy hour at that time constitute a small proportion of traffic sensitive network facilities.

[31] No, or almost no, weekend traffic will impose any capacity costs.

[32] Prices that varied optimally by time would smooth and spread traffic peaks, increasing the proportion of traffic that pressed on capacity and decreasing the proportion of traffic that should be charged a price of zero. Prices that vary hour by hour, however, are very unlikely to be practical. The closest feasible approximation is likely to be uniform pricing throughout some peak period. Such pricing will depress all calling within the period, and will not achieve the peak smoothing of optimally varied prices. Thus with any feasible set of pricing, the traffic profile is likely to continue to have a peaked busy hour, and it likely will continue to be true that most traffic will impose only minimal capacity costs.

majority of all traffic, but departs further from the optimal signal than does a uniform price during the times when usage imposes capacity costs on the terminating carrier. In the absence of detailed cost and demand information, no clear-cut conclusion is possible about which pricing structure, on balance, sends more efficient (or less inefficient) pricing signals.

4. Bill and Keep versus Peak/Off Peak Usage Pricing

Usage charges do not have to be uniform, and a standard response to the problem of recovering capacity costs with usage charges, both in theory and in the practice of telephone pricing, has been to set higher charges for peak than off-peak usage. Can the inefficiencies of uniform prices be overcome by setting peak and off-peak rates?

Theoretical studies of optimal peak/off-peak pricing have assumed a particular pattern of demand for usage: uniform demand within each pricing period, with demand at a high and uniform plateau during the peak period, and a uniform but lower plateau during an off-peak period. This pattern makes it optimal to set just two price levels (with the off-peak price usually at zero). But in practice the pattern of telephone usage varies from hour to hour. A quick look back at the optimal pricing structure derived earlier suggests that such two-period peak/off-peak structures also fall short of optimality. Peak/off-peak structures typically identify just two or three rate periods, and charge uniform rates within each. In contrast, in the example of an optimal rate structure above, several different non-zero rates were necessary in order to smooth the traffic peak. For the usage patterns typical of telephone traffic, there is no peak period during which a uniform, relatively high price would be charged.

Setting theoretically optimal prices that differ from hour to hour will not be feasible in practice.[33] It will be difficult and costly to collect the detailed demand information necessary to calculate such prices, demand may be constantly shifting and require frequent changes in peak pricing periods, and it is costly to collect charges based

---

[33] The concept of feasible prices is discussed in Park and Mitchell, *op. cit.*, pp. 5-6.

on such prices.[34] Furthermore, consumers likely would find it difficult to deal with such complicated pricing structures (assuming they were reflected in retail pricing). Varying prices would be unlikely to have the desired effect on consumer calling, even if implemented, because consumers are unlikely to understand and know the varying prices of calling at various times. In practice, only pricing structures that are feasible can (or should) be implemented. Simple peak/off-peak pricing with two or three pricing periods is feasible. Like uniform prices and bill and keep, however, simple peak/off-peak prices do not send fully optimal price signals. The question again is, how far do they depart from optimality? To clarify the exposition, we discuss only the case of two pricing periods.

Off-peak prices are easily evaluated -- assuming that the off-peak period is set so that no additional traffic during this period imposes capacity costs. Off-peak prices set at zero will be optimal, just as bill and keep sends optimal price signals for this traffic. If off-peak prices are not zero, they should still be lower than a uniform price, in which case they will impose smaller deadweight losses than the uniform price, but greater deadweight losses than the zero price of bill and keep.

The effect of the peak period rate is more complicated. Peak periods typically are relatively long; often, for example, they cover regular business hours or more. Such periods certainly will be longer than the system busy hour of the terminating carrier. Some facilities used by terminating traffic will have busy hours outside the system busy hour, but a long peak period almost surely will extend over periods when additional traffic imposes no capacity costs. Applying the peak period rate to this traffic generates a

[34] In fact, optimal pricing that fully took into account the variability in demand and cost would be even more complicated. Demand varies systematically not just by hour of the business day, but by day of the week and time of the year. Furthermore, the level of demand does not shift sharply when the hour is struck, but varies continuously across time. As suggested above, different network facilities will face congestion at different times and facility cost will vary by location. Fully optimal pricing in principle would take this into account, varying price not only with time of day but with the routing of the call.

deadweight loss. Furthermore, the peak period price generates a larger deadweight loss for this traffic than the uniform price because it is higher.[35]

The peak period price also is likely to be too low during some of all of the portion of the peak period when additional traffic does impose capacity costs. Because the peak period still includes traffic that does not impose capacity costs, the calculated price will be lower than optimal for some or all traffic that does impose capacity costs. In effect, pricing in the peak replicates the pattern of inefficiencies of uniform pricing. Peak period prices may be right "on average" over the period, but will be too low for some traffic, too high for most of the rest of the traffic, and just right only by accident.

A peak/off peak price structure should send more efficient pricing signals than uniform prices (so long as both generate the same total revenue).[36] It still is not possible, however, to reach any general conclusion about the relative efficiency of pricing signals from peak/off peak usage pricing and bill and keep arrangements. As before, the ranking depends on detailed cost and demand information, and now on the design of peak/off peak pricing as well.

5. Level of Pricing

The discussion so far has focused only on effects of the structure of usage sensitive price. The assumption has been that the cost of capacity was known, and that usage prices do no more than recover the capacity costs imposed by terminating traffic. In other words, the implicit assumption has been that the overall level of usage based prices was correct, and the only issue was the effect of the structure of those prices. In

---

[35] The optimal time-varying pricing structure described above also charged non-zero prices outside the busy hour. These prices, however, varied depending on demand, order to smooth off the peak of the traffic distribution. Prices varied so that demand was not suppressed by more than was necessary to smooth the traffic distribution. A uniform peak period price will tend to suppress demand more than is necessary during some hours of the peak period outside the busiest hour, and this generates a deadweight loss.

[36] This assumes that the peak and off peak periods are not set perversely, for example by setting a peak period that does not include the busy hour. In principle it should always be possible to do at least as well as with uniform prices, since uniform pricing can be replicated by setting the same rate in the peak and off-peak period.

fact, these costs may not be known, or prices may not be set at the most efficient level.[37] If, for whatever reason, the level of usage sensitive prices is set too high, that will be an additional source of inefficiency. In particular, there is general agreement that interstate switched access charges are set well above costs. Setting usage prices at this level surely would impose additional efficiency losses. (As discussed below, setting high interconnection prices is also likely to deter the development of competition and impose losses of dynamic efficiency.)

VI. The Effect of the Compensation Arrangements on Transactions Costs

The efficiency of price signals is not the only criterion for evaluating the efficiency of compensation arrangements or choosing among them. A second criterion is the effect of compensation arrangements on cost.

A. Tradeoffs: Triangles and Rectangles

In economic theory, overall efficiency depends on satisfying a number of conditions. Having the appropriate price signal -- in textbook theory setting price equal to marginal cost -- is only one of these conditions. A second condition necessary to achieve efficiency is to produce in the most efficient possible way and to minimize cost for any given level of output. Ideally, one both minimizes cost and obtains efficient pricing signals.

Sometimes, however, that may not be possible. It may be costly to get sufficient cost information to price "perfectly." Or it may be costly to monitor usage and collect revenue. One then has to tradeoff the effects on efficiency of a better price signal but higher costs, against a less accurate price signal but lower costs.

---

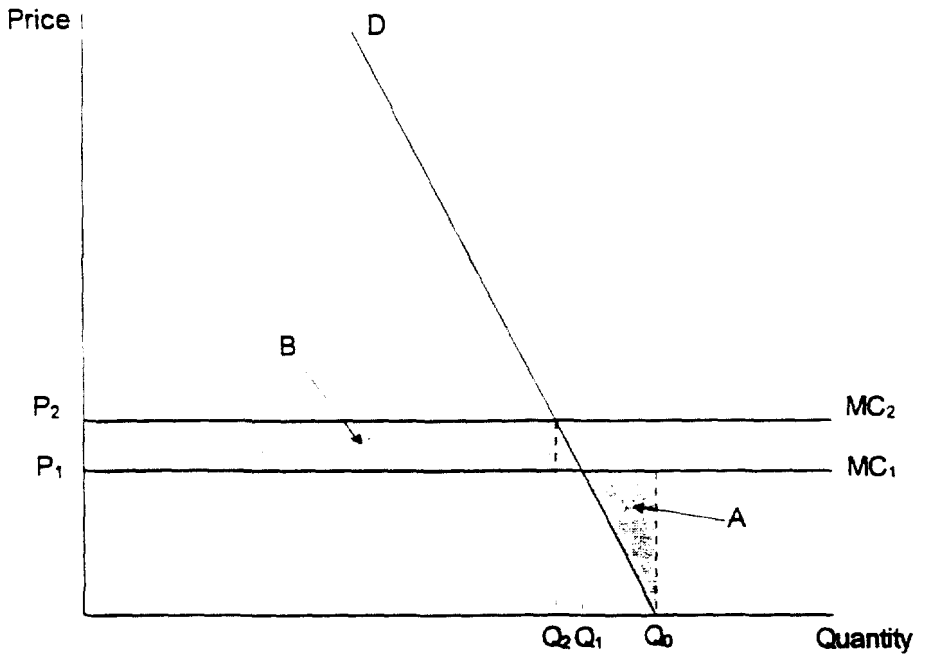
[37] The question of the efficient level of pricing raises many complicated issues whose discussion is beyond the scope of this paper. Among these are questions of whether prices should be set to recover long or short run costs, marginal or total service incremental costs, and whether markups above (some measure of) cost are appropriate, and if so what are appropriate justifications for such markups. For discussions of some of these issues see Bridger M. Mitchell, Werner Neu, Karl-Heinz Neumann and Ingo Vogelsang, "The Regulation of Pricing of Interconnection Services" in G. R. Brock, ed., *Toward a Competitive Telecommunications Industry*, Lawrence Erlbaum Assoc, 1995.

Figure 3 illustrates such a tradeoff with a simple graph of demand and cost. D is the demand curve, showing the amount consumers would purchase at each price. MC_1 plots the constant marginal cost of additional output, absent any costs of charging and collecting a price. The optimal solution, if it were possible, would be to charge P_1, where price equals marginal cost, in which case consumers would buy (and producers would produce) the quantity Q_1. Assume, now, that collecting a price for each unit consumed will increase marginal cost, because each unit sold must be recorded and revenue collected. If a price is charged, marginal cost increases to MC_2. Now the combination of P_1 and Q_1 is unachievable. The choice is between charging a price of zero with consumption of Q_0, and setting price at P_2, (equal to the higher MC_2 with non-zero prices), in which case quantity consumed declines to Q_2. The combination of a price of zero and Q_0 generates a deadweight loss shown by the shaded triangle A.[38] This represents the difference between how much consumers value the additional units of output, given by D, and the larger amount it costs to produce them, given by MC_1. If instead a price is charged, the efficiency cost is the increase in cost for each unit of output, which is shown by the shaded rectangle B.[39] Overall, failing to charge a price for output will be more efficient if the deadweight loss of triangle A is smaller than the rectangle B.

[38] To simplify, we assume that the producer has other ways of collecting revenue to cover the costs of producing Q_0. For example, the "product" here might be packets of ketchup at McDonald's, which can charge a separate price for each ketchup packet, or cover the cost by charging slightly more for hamburgers. The slightly higher price for hamburgers in this example would generate an additional deadweight loss that should be added to that shown in Figure 2 to calculate the full effect of giving away ketchup packets.

[39] The rectangle represents the increased cost per unit of output, $MC_2 - MC_1$, times the output of Q_2.

Figure 3
Efficiency Trade-off: Deadweight Loss Versus Higher Transaction Costs



B. Costs with Bill and Keep and Usage Sensitive Pricing

Under bill and keep, neither the LEC nor CMRS provider needs to track or bill for the amount of interconnected traffic it receives from the other carrier. These functions, and their costs, are necessary with usage sensitive pricing.

Usage sensitive compensation arrangement will unquestionably impose higher transactions costs than bill and keep arrangements, although it is not clear that incurring these costs will necessarily lead to more efficient pricing signals. The analysis of the previous section showed that neither usage sensitive pricing nor bill and keep is able to send fully optimal pricing signals. Nor does the analysis support a general conclusion that usage sensitive pricing will necessarily send better, more efficient pricing signals than bill and keep compensation arrangements, as the efficiency of the signals sent by usage sensitive pricing will vary with the structure and level of those prices. In either case, the higher costs of administering usage sensitive pricing are a factor counting in

favor of bill and keep arrangements, either as an offset to somewhat less efficient pricing signals with bill and keep, or as a factor augmenting the bill and keep's more efficient pricing signals.

Survey responses from CTIA members identified various types of costs they would save with bill and keep compensation arrangements. If they no longer had to make usage sensitive payments for traffic sent to LECs, they would save costs of administrative and financial personnel and supporting services necessary to audit, reconcile, verify, and pay bills. One system indicated that it employed two full time clerks to handle LEC billing, and another that one full time staff member was devoted to analyzing bills from the LEC. Most cellular systems answering the questionnaire are not now paid for the termination of LEC-originated traffic. If compensation for these costs were based on usage payments rather than bill and keep arrangements, respondents indicated they would incur personnel and other costs to collect the necessary data, to prepare bills, to handle accounts receivable and payable, and to manage the process. Several systems also reported that they do not now have the ability to measure traffic received from LECs, and that adding this capability would involve a significant expense.

Finally, it is worth noting that some of the transactions costs must be committed up-front to implement usage sensitive pricing. Such costs likely include the costs of regulatory proceedings to collect cost information and set rates, the cost to providers of establishing procedures, developing software, and training personnel to implement the pricing, and the costs of any special equipment that must be installed to measure usage. Once these costs are incurred, there is no way to go back and undo them, or reduce their burden, if usage sensitive pricing proves suboptimal and a shift is made to bill and keep or some other compensation arrangement.

VII. Effects on Competition and Dynamic Efficiency

We have analyzed how compensation arrangements, by influencing pricing signals and transaction costs, affect static efficiency. The analysis of pricing signals asks how pricing affects consumers' usage of a given set of services and suppliers. The analysis of transactions costs asks how costs of given services and suppliers are affected